IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 27 August 2023, accepted 5 September 2023, date of publication 8 September 2023,
date of current version 15 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3313432

RESEARCH ARTICLE

Monocular 3D Object Detection From Comprehensive Feature Distillation Pseudo-LiDAR

CHENTAO SUN[1], CHENGRUI XU[1], WENXIAO FANG[2], AND KUNYUAN XU[1]
[1]School of Physics, South China Normal University, Guangzhou 510006, China
[2]School of Integrated Circuit, Sun Yat-sen University, Shenzhen 528406, China

Corresponding authors: Kunyuan Xu (xuky@scnu.edu.cn) and Wenxiao Fang (fangwenx@mail.sysu.edu.cn)

This work was supported by the Science and Technology of Guangzhou under Grant 2019050001.

ABSTRACT The use of knowledge distillation in monocular 3D object detection has been explored by incorporating a LiDAR model as the teacher network to transfer knowledge to a monocular network. However, LiDAR data and images belong to distinct data types, and their respective models exhibit significant structural disparities. These differences serve as constraints to the complete and comprehensive transmission of depth information from the teacher network to the student network. To overcome these limitations, we propose an end-to-end network with Comprehensive Feature Knowledge Distillation (CFKD) monocular pseudo-LiDAR. This method transforms monocular images into pseudo-LiDAR and feeds them into a student LiDAR network which receives distilled knowledge from a teacher LiDAR network. By leveraging the similarity in the network structures of the teacher and student LiDAR networks, our approach efficiently utilizes the LiDAR information via comprehensive distillation of features. We assessed our method's efficient implementation on the kitti3D dataset. Our methods achieved an improvement of 4.67 for $\mathrm{AP}_{\mathrm{BEV}}$ in the moderate category and 2.65 for $\mathrm{AP}_{\mathrm{BEV}}$ in the hard category on the test set.

INDEX TERMS CFKD, knowledge distillation, monocular 3D object detection, pseudo-LiDAR, autonomous driving.

I. INTRODUCTION

In the realm of autonomous driving technology [1], 3D object detection plays a crucial role. LiDAR technology [2], [3], [4], which employs computer vision and deep learning, is commonly used for this purpose. However, the expense of LiDAR sensors and the sparse nature of the generated LiDAR limit the data available. Consequently, image-based 3D detection methods have become popular due to their cost-effectiveness. The primary difficulty with image-based detection is obtaining accurate depth information. Such methods comprise stereo image-based detection [5], [6], [7] and monocular image-based detection [8], [26]. In stereo image-based detection, the input consists of images captured by a left and a right camera with the same frequency. The stereo detection network then matches the left and right images to estimate the depth of each pixel. From this data, the three-dimensional position of the object is determined. Nonetheless, stereo detection's reliance on two synchronized cameras limit its applicability owing to computational complexity, occlusion issues, and lighting conditions. In contrast, monocular 3D object detection solely relies on a single camera, offering a cost-effective solution that significantly reduces algorithm complexity relative to stereo detection.

Monocular 3D object detection [38], [39] relies on single images to estimate the 3D position of objects, using deep learning technology. However, such an approach faces inherent challenges, mainly due to the absence of intrinsic depth information within single images. In recent years, researchers have proposed monocular 3D object detection methods utilizing geometric prior constraints [7], which rely on the fact that real-world objects usually exhibit specific 3D shapes. This allows for the estimation of object depth through the ratio between image and real-world height. Another approach to obtaining 3D information from images is through

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

pseudo-LiDAR based 3D detectors [11], [12]. This technique uses a depth estimator [13], [14] to estimate a depth map of the input image, which is then transformed into a pseudo-LiDAR point cloud. Finally, the converted data is used for 3D object detection with a LiDAR-based detector. Recently, researchers have started employing knowledge distillation techniques in image-based 3D object detection [15], [16]. This method leverages the depth information present in LiDAR point clouds, transferring it to monocular 3D object detectors to enhance their accuracy and robustness.

It is essential to note that transferring knowledge from a LiDAR model to an image model is a complex process due to the substantial differences between their network structures and the types of data they handle. Consequently, the direct application of knowledge distillation in this scenario is not feasible. As a result, the previous attempts to distill information from LiDAR data into monocular images were not entirely successful. To address this limitation, this paper proposes an end-to-end network with Comprehensive Feature Knowledge Distillation (CFKD) monocular pseudo-LiDAR for 3D object detection. Specifically, this study adopts an end-to-end image-based detector [9] from monocular pseudo-LiDAR as the student network, while the LiDAR-based detector [3] utilized in the student network serves as the teacher model for knowledge distillation, which is integrated at corresponding positions within the student network. Since the teacher network takes real LiDAR point cloud data as input, and the pseudo-LiDAR data is used as input for the student network's LiDAR-based detector. They are both LiDAR modal information, eliminating any semantic gap between point cloud and image modalities.

By using the same detector for knowledge distillation, we can confidently transfer knowledge from different depths of the teacher network to the student network. This enables comprehensive feature distillation, which facilitates the comprehensive transmission of knowledge from the teacher model to the student model. Our paper consists of five sections. The first section is the Introduction. The second section is the Related Work, which is further divided into four parts: Knowledge Distillation, Pseudo-LiDAR based 3D Detection, LiDAR-based 3D Object Detection, and the difference between our framework and predecessors. The third section is the Method, which consists of three parts: Framework overview, Change of Representation, and Loss. The fourth section is the Experiments, divided into five parts: Dataset, Implementation Details, Results on KITTI dataset, Ablation Studies, and Qualitative Results. The fifth section is the Conclusion.

In summary, this article makes the following contributions:

•We employed a two-fold approach to utilize the LiDAR data, leveraging it for both depth estimation and 3D object detection tasks.

•We propose a comprehensive feature knowledge distillation method based on an end-to-end modality-transfer network. This fully leverages the LiDAR information in the teacher network.

•Our framework was evaluated on the KITTI 3D dataset. In monocular 3D object detection, we achieved the best results based on the average values of three different levels of occlusion: easy, moderate, and hard. These findings highlight the effectiveness and capabilities of our proposed framework.

II. RELATED WORKS

A. KNOWLEDGE DISTILLATION

KD (Knowledge Distillation) [17] is a technique used for compressing and optimizing neural network models by transferring knowledge from a complex and information-rich model (the "teacher model") to a simpler and less informative model (the "student model"). The teacher model provides additional guidance to the student model during the process, improving the inference speed and accuracy of the student model, while reducing the storage and computational cost of the model. Typically, the teacher and student models have the same inputs, but the student model has fewer layers or parameters. In this work, our teacher model is a LiDAR model that takes real LiDAR point clouds as input and performs comprehensive feature distillation on the pseudo-LiDAR model.

B. PSEUDO-LIDAR BASED 3D DETECTION

Pseudo-LiDAR based 3D detectors [12], [18], [19] use images as input and a depth estimator to generate depth information which is converted into pseudo-LiDAR and fed into a high-performance LiDAR detector to estimate 3D positions. Recently, [11] used monocular images as input and a monocular depth estimator for end-to-end monocular pseudo-LiDAR 3D object detection, incorporating a pose estimation network that leveraged both the previous and current frames of the input images. However, our study focuses on knowledge distillation using only monocular images as input, without incorporating a pose estimation network.

C. LIDAR-BASED 3D OBJECT DETECTION

In recent years, there has been significant progress in the development of 3D object detection algorithms utilizing LiDAR technology. These algorithms are designed to process unordered point cloud data collected by LiDAR sensors, allowing for the estimation of objects' 3D positions based on this information. LiDAR-based 3D object detection approaches [34], [35], [36] comprise a range of methods, including point-based approaches such as PointNet [20] and PointNet++ [21]. These techniques make use of LiDAR data as input and employ network frameworks to extract point-level features for estimating objects' 3D positions. Alternatively, there are voxel-based methods [22] that involve transforming LiDAR point clouds into voxel features. These voxel-based representations are then fed into the network for estimating the 3D position of objects. This approach can be likened to processing pixels in an image, with voxels serving as the 3D counterparts of these pixels.

D. THE DIFFERENCE BETWEEN OUR FRAMEWORK AND PREDECESSORS

Several attempts have been made to apply knowledge distillation to monocular 3D object detection. For example, MonoDistill [16] utilizes knowledge distillation by translating LiDAR modal representation into image modal information. Similarly, CMKD [15] involves feature distillation by converting image modal representations into Bird's Eye View (BEV) features. However, it only conducts a single layer of BEV feature distillation, and it can only use voxel-based LiDAR networks to distill the student network, indicating the limitations of its framework. In contrast, our proposed framework not only enables comprehensive and thorough feature distillation, but also has greater versatility. Our framework is applicable to all types of LiDAR models for distillation. LPCG [23] applies a LiDAR-based detector to produce pseudo labels that guide the image-based detector. In contrast, DA-3d [24] follows a non-end-to-end training strategy, utilizing a fixed-depth estimator and a 2D detector, with adaptation limited to the feature extractor in the feature domain. Nonetheless, the framework of DA-3d is two-stage, and its detection performance is limited by the accuracy of the monocular depth estimator.

In contrast, our framework takes an end-to-end approach, which allows for the propagation of distillation information and label information back to the depth estimator during the training process. Previous methods in monocular 3D object detection did not fully distill LiDAR modal information into monocular 3D detectors, and only distilled feature detector outputs. Our framework overcomes these limitations through an end-to-end approach that utilizes a LiDAR detector as the teacher model for feature distillation using real LiDAR point clouds as input. As the LiDAR detector is same, our framework can distill all layers of features in the model, leveraging information from real LiDAR modalities to enhance monocular 3D detection's effectiveness. During the validation and testing stages, our approach solely utilizes monocular images as input without the use of additional inputs.

III. METHOD

A. FRAMEWORK OVERVIEW

Fig. 1 depicts an overview of our proposed end-to-end knowledge distillation pseudo-LiDAR framework for object detection. Initially, a monocular image is fed into a monocular depth estimator to estimate the depth map. The estimated depth map is then compared with the ground truth obtained by projecting the real LiDAR. Next, the depth map is transformed into a pseudo-LiDAR, which is in turn inputted into a soft quantization PIXOR [18] for 3D object detection. During the 3D object detection stage, we use a pre-trained PIXOR as the teacher model with the real LiDAR point cloud as input, and conduct knowledge distillation on the student model. Because the teacher model obtains richer and more precise depth information from the real LiDAR, we can ensure the feasibility of our method. Moreover, our network framework is end-to-end, and the knowledge obtained from the LiDAR detection part of the student network can be back-propagated to the depth estimator.

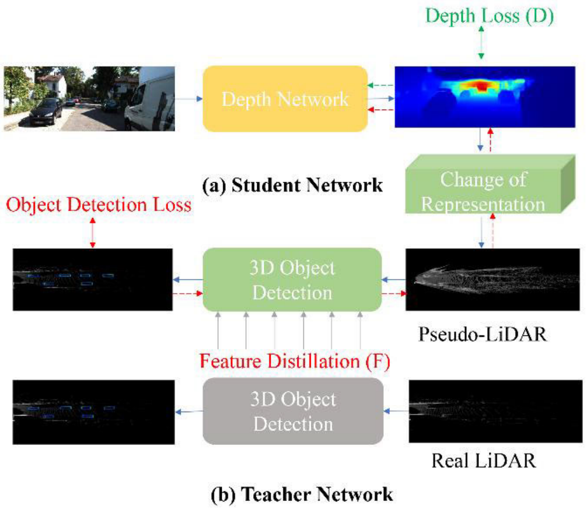


FIGURE 1. **Overview of our distillation framework. Forward propagation is indicated by solid blue arrows, whereas backward propagation is indicated by dotted arrows. Firstly, we input monocular images into a depth estimator to obtain the depth map. Next, the depth map is converted to a pseudo-LiDAR to obtain detection results through a 3D detector. During the 3D detection stage, we conducted knowledge distillation using a similar 3D detector as the teacher network, which received input from a real LiDAR.**

For the training phase, we provide both monocular images and their matching LiDAR point clouds. The student model processes only monocular images, while the teacher mode receives real LiDAR point clouds. We only apply knowledge distillation to the 3D detection component of the student network. As the 3D detectors in both the teacher and student models have the same structure, we can comprehensively apply knowledge distillation to different deep features in the student model's 3D detector. During the inference phase, we only use monocular images for end-to-end evaluation and testing.

B. CHANGE OF REPRESENTATION

1) PSEUDO-LiDAR POINT CLOUD

The module for estimating depth predicts the depth $Z(u, v)$ for each pixel (u, v) in the image. This depth is then transformed into a pseudo-LiDAR point cloud (x, y, z).

$$z = Z(u, v), x = (u - c_U) \cdot /f_U, y = (v - c_V) \cdot /f_V, \qquad (1)$$

Within the equation, (c_U, c_V) indicates the camera center, while f_U and f_V represent the horizontal and vertical focal lengths. On the KITTI benchmark, monocular pseudo-LiDAR detection has now attained cutting-edge performance, and our framework is based on this significant advancement.

2) SOFT QUANTIZATION

To prepare the point cloud for network processing, we utilize quantization to voxelize it into a 3D tensor. Previously, pseudo-LiDAR detection took place in two separate stages. For backward propagation of information to the depth estimator, we utilize soft quantization, a type of change of representation (CoR), which is explained in reference [18]. The soft quantization method defines a three-dimensional occupancy tensor T with M bins, expressed as follows:

$$T\left(m\right)=T\left(m,m\right)+\frac{1}{|N_m|}\sum_{m'\epsilon N_m}T\left(m,m'\right), \tag{2}$$

In the equation, $m\in$ *{1, …, M}* is a bin in T, N_m denotes the set of neighboring boxes of m and

$$T\left(m,m'\right)=\begin{cases}0 & if\,|P_{m'}|=0,\\ \frac{1}{|P_{m'}|}\sum_{p\epsilon P_{m'}}e^{-\frac{||p-\hat{p}_m||^2}{\sigma^2}} & if\;|P_{m'}|>0.\end{cases} \tag{3}$$

where $\hat{p}_m$ denotes the center of bin m. We define P_m as the set of points within bin m, which can be mathematically represented as $P_m=\{p\in P,s.t.m=||p-\hat{p}'_m||^2\}$.

C. LOSS

The elaborate distillation structure of our framework is displayed in Fig. 2. Specifically, both networks depicted in the figure are implemented as soft quantized PIXOR models [18] to ensure precise alignment of intermediate features between the student and teacher networks. we distilled the 9 features of the network, which enabled us to completely transfer the information from the teacher network to the student network. We employed smooth *L1* loss for this distillation process,

$$L_{kd}=\text{SmoothL1Loss}\left(F_t,F_s\right)$$

where F_t represents the output features of the intermediate layer in the teacher network, and F_s represents the output features of the intermediate layer in the student network. Additionally, we employed the smooth *L1* loss to compare the estimated depth map with the ground truth. Projecting the LiDAR point cloud onto the image plane allowed for the acquisition of the ground truth.

$$L_{depth}=\text{SmoothL1loss}\left(Z(u,v),Z*(u,v)\right)$$

where *Z∗(u, v)* is the ground truth, and *Z (u, v)* is the predicted depth.

Our 3D detection loss consists of classification loss and regression loss. While smooth L1 loss is still used to calculate the regression loss, Binary CrossEntropy loss is used to calculate the classification loss. The 3D detection loss can be expressed as follows:

$$L_{det}=L_{cls}+L_{reg}$$

Our total loss is as follows, where the value of λ is 0.1.

$$L=\lambda\,L_{det}+L_{depth}+L_{kd}$$

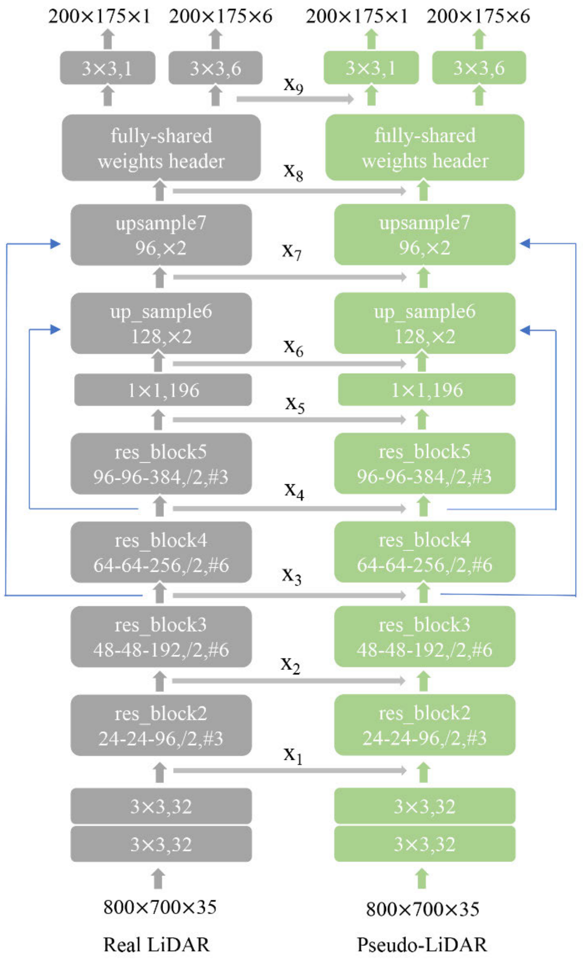

200×175×1
200×175×6
200×175×1
200×175×6
3×3,1
3×3,6
x9
3×3,1
3×3,6
fully-shared weights header
x8
fully-shared weights header
upsample7 96,×2
x7
upsample7 96,×2
up_sample6 128,×2
x6
up_sample6 128,×2
1×1,196
x5
1×1,196
res_block5 96-96-384,/2,#3
x4
res_block5 96-96-384,/2,#3
res_block4 64-64-256,/2,#6
x3
res_block4 64-64-256,/2,#6
res_block3 48-48-192,/2,#6
x2
res_block3 48-48-192,/2,#6
res_block2 24-24-96,/2,#3
x1
res_block2 24-24-96,/2,#3
3×3,32
3×3,32
3×3,32
3×3,32
800×700×35
800×700×35
Real LiDAR
Pseudo-LiDAR

FIGURE 2. **A detailed description of the distillation part of the frame. Here on the left is the teacher network, on the right is the LiDAR network part of the student network, and x represents the features of the network output.**

IV. EXPERIMENTS

A. DATASETS

The KITTI 3D dataset served as the basis for our experimental evaluation [1]. For training, validation, and testing, the dataset includes 3712, 3769, and 7518 images, respectively. The dataset also contains the associated 64-beam LiDAR data. Labels are annotated on the training and validation sets. We use the APBEV measure, which stands for the average precision in bird's-eye view, to assess the performance of our suggested strategy on both the validation and test sets in terms of average precision (AP). Using categories like easy, moderate, and hard to describe different occlusion degrees, the KITTI dataset classifies items according to the degree of occlusion.

B. IMPLEMENTATION DETAILS

We utilized the baseline model proposed by [9], with a modification that only monocular images were fed into the model, and its pose network was not utilized. Specifically, our depth

TABLE 1. Result for Car on KITTI validation set. We compared our method with previous results and reported the AP_{BEV}. We use bold font to represent the best results and underline to indicate the second-best results.

Method	AP_{BEV}, IoU=0.5			AP_{BEV}, IoU=0.7		
	Easy	Moderate	Hard	Easy	Moderate	Hard
PL-MONO [11]	70.8	49.4	42.7	40.6	26.3	22.9
DA-3D [24]	-	-	-	<u>45.5</u>	<u>32.7</u>	27.1
GUPNet [25]	61.78	47.06	40.88	31.07	22.94	19.75
MonoDistill [16]	<u>71.45</u>	53.11	46.94	33.09	25.40	22.16
MonoPGC [31]	-	-	-	34.06	24.26	20.78
ADD [32]	-	-	-	40.38	29.07	25.05
MonoATT [33]	-	-	-	38.93	29.76	25.73
Baseline [9]	70.86	<u>60.23</u>	<u>54.54</u>	37.42	31.05	<u>29.13</u>
Ours	81.44	72.35	65.02	52.82	45.94	40.07

TABLE 2. Result for Car on KITTI test set. We compared our method with previous results and reported the AP_{BEV} at IoU=0.7. We use bold font to represent the best results and underline to indicate the second-best results.

Method	Auxiliary Data	AP_{BEV}			
		Easy	Moderate	Hard	Average
M3D-PRN [10]	None	21.02	13.67	10.23	14.97
PatchNet [27]	LiDAR	22.97	16.86	14.97	18.24
D4LCN [28]	LiDAR	22.51	16.02	12.55	17.03
CaDDN [29]	LiDAR	27.94	18.91	17.19	21.35
GUPNet [25]	None	30.29	21.19	18.20	23.23
MonoCon [30]	None	31.12	22.10	19.00	24.07
MonoDistill [16]	LiDAR	31.87	22.59	19.72	24.72
CMKD [15]	LiDAR	33.69	23.10	20.67	25.82
MonoPGC [31]	LiDAR	32.50	23.14	20.30	25.31
ADD [32]	LiDAR	<u>35.20</u>	23.58	20.08	26.28
MonoATT [33]	None	**36.84**	<u>24.42</u>	<u>21.88</u>	<u>27.71</u>
Ours	LiDAR	34.28	**29.09**	**24.53**	**29.30**
Improvement		-2.56	+4.67	+2.65	+1.59

estimator utilizes Monodepth2 [37]. The LiDAR detection component of the student network and the teacher network utilize soft quantized PIXOR models. Our framework was implemented on PyTorch, with most of the hyperparameters following [9]. The RMSprop optimizer was utilized, and the model was trained for 25 epochs with a learning rate of 5e-5, which was decreased by a factor of ten at the beginning of each 10 epochs. One was chosen as the batch size. Pre-trained weights were used to initialize the depth estimator. The teacher network performed distillation on the student network utilizing the pre-trained PIXOR, without engaging in backpropagation.

C. RESULTS ON KITTI DATASET

The KITTI validation set and test set were both used to evaluate the results of our research. Table 1 and Table 2, respectively, show the outcomes of the experiment. We uploaded our test outcomes to the KITTI website for review since the KITTI test set lacks ground truth labels. On the test sets, our method showed improvements of 4.67 in the moderate category, 2.65 in the hard category, and an average improvement of 1.59 across all three categories. On the validation sets, our technique significantly improved three categories, as can be seen in the table 1. Moreover, in every category, our method produced cutting-edge outcomes. These improvements demonstrate the effectiveness of our method. Especially when compared to the baseline model on the validation set. For each category, we have achieved an improvement of over ten points, highlighting the powerful impact of our distillation method.

D. ABLATION STUDIES

We conducted ablation studies to examine the performance of our approach using both deep feature distillation and shallow feature distillation. We present the results of these experiments in Table 3, where (a) refers to our baseline model. In (b), we distilled the features in the front part of the 3D detector, while (c) aimed to investigate the impact of feature distillation in the latter part. Lastly, (d) involved distilling all features. A comparison of (a) with (b) highlights that our feature distillation method is successful. Furthermore, we observed that (c) outperforms (b), since (b) exclusively distilled shallow features from the 3D detection network, which only captures some texture information about the objects. In contrast, (c) distills deep features from the latter part of the network that convey the overall information of the objects. As a result of the fusion of certain shallow features into the latter part of the features, it includes not only the overall features of the object but also the texture features of the object. Hence, teacher network is able to instruct our student network more, leading to better performance of (c) compared to (b). As for the (d) experiment, it produced the best results. This is because it not only transfers knowledge from shallow features but also utilizes deep features. This allows for correcting features at different depths in the student network, thereby avoiding detection biases created by inaccuracies in the pseudo-LiDAR. Consequently, the network achieves superior performance.

TABLE 3. Ablation studies on the KITTI validation set, x_1 to x_9 represent features at different depths within the network.

Group	Feature Distillation									AP$_{BEV}$, IoU=0.5			AP$_{BEV}$, IoU=0.7		
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Easy	Moderate	Hard	Easy	Moderate	Hard
(a)										70.86	60.23	54.54	37.42	31.05	29.13
(b)	√	√	√	√	√					78.06	67.88	61.36	49.14	39.71	34.38
(c)						√	√	√	√	78.14	71.86	63.80	48.93	41.19	35.27
(d)	√	√	√	√	√	√	√	√	√	81.44	72.35	65.02	52.82	45.94	40.07

TABLE 4. Ablation studies on the KITTI validation set, "D" represents the deep loss, while "F" represents our feature distillation loss.

D	F	AP$_{BEV}$, IoU=0.5			AP$_{BEV}$, IoU=0.7		
		Easy	Moderate	Hard	Easy	Moderate	Hard
		64.35	50.39	44.03	28.25	23.19	21.01
√		70.86	60.23	54.54	37.42	31.05	29.13
	√	80.00	72.30	64.21	48.53	42.45	38.84
√	√	81.44	72.35	65.02	52.82	45.94	40.07

TABLE 5. The ablation studies conducted on the KITTI validation set resulted using different feature distillation losses.

Feature Loss	AP$_{BEV}$, IoU=0.5			AP$_{BEV}$, IoU=0.7		
	Easy	Moderate	Hard	Easy	Moderate	Hard
L1	79.34	69.95	63.24	46.71	40.05	34.35
L2	78.33	67.64	63.69	48.54	41.66	36.31
SmoothL1loss	81.44	72.35	65.02	52.82	45.94	40.07

Table 4 presents our ablation studies exploring the use of deep loss and feature distillation loss independently. It is apparent that utilizing either of the losses individually results in improved network performance. Specifically, applying only the feature distillation loss significantly enhances the results.

Nonetheless, the network attains the best performance when both losses are used together. This demonstrates that our network effectively leverages the information from the LiDAR data, resulting in enhanced detection results.

Table 5 displays our investigation on the influence of various feature distillation losses on the distillation effect. We performed experiments with three types of loss functions. Based on the information presented in the table, it is evident that employing L2 loss as the distillation loss produces superior outcomes in comparison to utilizing L1 loss. Using smooth L1 loss yields the most optimal results.

E. QUALITATIVE RESULTS

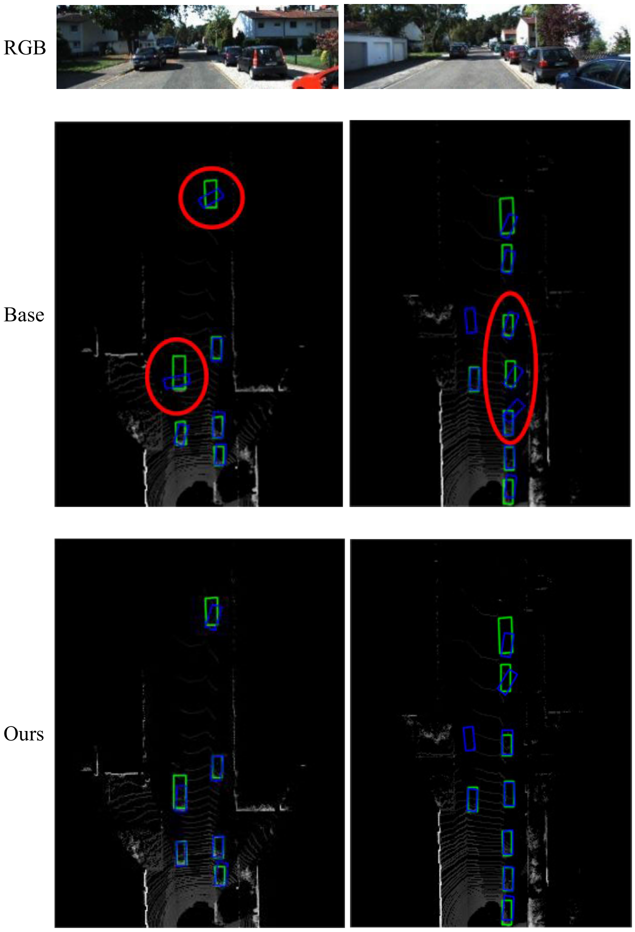


FIGURE 3. A Comparison of 3D object detection results between the baseline model and our proposed method. The green color indicates the label, while the blue color indicates the detection result.

To visually illustrate the superior performance of our model, Fig. 3 conducted a visualization comparing the detection results of our model with those of the baseline model on the bird's-eye view. The first row of images represents the input images, while the second row displays the detection results of the baseline model. The third row exhibits the detection results obtained from our method. The first column corresponds to the first input image and its corresponding detection results, whereas the second column represents the second input image and its detection results. The green boxes represent the ground truth, while the blue boxes represent the detection results. We use red circles to highlight the incorrectly detected results of the baseline model. In the first column, we can see that the baseline model has inaccuracies in both localization and orientation, while our method achieves more accurate localization and orientation detection. As for the second column, we can see that the detection results circled by the baseline model are inaccurate in terms of direction and have extra boxes, while our method provides better detection results. These results confirm the reliability and effectiveness of our approach.

V. CONCLUSION

In this study, we propose an end-to-end comprehensive feature knowledge distillation 3D object detection from monocular pseudo-LiDAR, aiming to maximize the utilization of depth information from real LiDAR and effectively transfer it to monocular 3D object detector. Additionally, we explore distilling knowledge from both deep and shallow features to investigate the effectiveness of transferring knowledge from different depth features. Although our approach has made significant progress, our method did not achieve optimal performance in terms of AP_{BEV} values for the easy category on the KITTI test set and our network structure is comparatively intricate and time-consuming. Our approach involves the conversion from image modality to LiDAR modality. Due to the difficulty in converting from images to LiDAR data, the uncertainty of the network is increased. To address this, future research can explore the potential of same-modal data distillation, such as extracting information from binocular data to improve monocular 3D detection.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[2] H. Fazlali, Y. Xu, Y. Ren, and B. Liu, "A versatile multi-view framework for LiDAR-based 3D object detection with guidance from panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17171–17180.

[3] B. Yang, W. J. Luo, and R. Urtasun, "PIXOR: Real-time 3D object detection from point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7652–7660.

[4] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection," *Int. J. Comput. Vis.*, vol. 131, pp. 531–551, Nov. 2022.

[5] X. Guo, S. Shi, X. Wang, and H. Li, "LIGA-Stereo: Learning LiDAR geometry aware representations for stereo-based 3D detector," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3133–3143.

[6] Y. L. Chen, S. J. Huang, S. Liu, B. Yu, and J. Jia, "DSGN++: Exploiting visual-spatial relation for stereo-based 3D detectors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4416–4429, Apr. 2023.

[7] J. Sun, L. Chen, Y. Xie, S. Zhang, Q. Jiang, X. Zhou, and H. Bao, "Disp R-CNN: Stereo 3D object detection via shape prior guided instance disparity estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10545–10554.

[8] K. Huang, T. Wu, H. Su, and W. H. Hsu, "MonoDTR: Monocular 3D object detection with depth-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4002–4011.

[9] C. Kim, U.-H. Kim, and J.-H. Kim, "Self-supervised 3D object detection from monocular pseudo-LiDAR," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2022, pp. 1–6.

[10] G. Brazil and X. Liu, "M3D-RPN: Monocular 3D region proposal network for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9286–9295.

[11] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8437–8445.

[12] X. Weng and K. Kitani, "Monocular 3D object detection with pseudo-LiDAR point cloud," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 857–866.

[13] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6602–6611.

[14] Y. Y. Li, Y. T. Chen, J. W. He, and Z. Zhang, "Densely constrained depth estimator for monocular 3D object detection," in *Computer Vision—ECCV 2022* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2022, pp. 718–734.

[15] Y. Hong, H. Dai, and Y. Ding, "Cross-modality knowledge distillation network for monocular 3D object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 87–104.

[16] Z. Chong, X. Ma, H. Zhang, Y. Yue, H. Li, Z. Wang, and W. Ouyang, "MonoDistill: Learning spatial features for monocular 3D object detection," 2022, *arXiv:2201.10830*.

[17] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[18] R. Qian, D. Garg, Y. Wang, Y. You, S. Belongie, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "End-to-end pseudo-LiDAR for image-based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5880–5889.

[19] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving," 2019, *arXiv:1906.06310*.

[20] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.

[21] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[22] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[23] L. Peng, F. Liu, Z. X. Yu, S. Yan, D. Deng, Z. Yang, H. Liu, and D. Cai, "LiDAR point cloud guided monocular 3D object detection," in *Computer Vision—ECCV 2022* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2022, pp. 123–139.

[24] X. Ye, L. Du, Y. Shi, Y. Li, X. Tan, J. Feng, E. Ding, and S. Wen, "Monocular 3D object detection via feature domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2020, pp. 17–34.

[25] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, and W. Ouyang, "Geometry uncertainty projection network for monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3091–3101.

[26] J.-Q. Yu and S.-C. Pei, "Perspective-aware convolution for monocular 3D object detection," 2023, *arXiv:2308.12938*.

[27] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, and W. Ouyang, "Rethinking pseudo-LiDAR representation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 311–327.

[28] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, and P. Luo, "Learning depth-guided convolutions for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 4306–4315.

[29] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun.*, Jun. 2021, pp. 8551–8560.

[30] X. Liu, N. Xue, and T. Wu, "Learning auxiliary monocular contexts helps monocular 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1810–1818.

[31] Z. Wu, Y. Gan, L. Wang, G. Chen, and J. Pu, "MonoPGC: Monocular 3D object detection with pixel geometry contexts," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 4842–4849.

[32] Z. Wu, Y. Wu, J. Pu, X. Li, and X. Wang, "Attention-based depth distillation with 3D-aware positional encoding for monocular 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 2892–2900.

[33] Y. Zhou, H. Zhu, Q. Liu, S. Chang, and M. Guo, ‘‘MonoATT: Online monocular 3D object detection with adaptive token transformer,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17493–17503.

[34] N. Lopac, I. Jurdana, A. Brnelić, and T. Krljan, ‘‘Application of laser systems for detection and ranging in the modern road transportation and maritime sector,’’ *Sensors*, vol. 22, no. 16, p. 5946, Aug. 2022.

[35] L. Wen and K. Jo, ‘‘Fast and accurate 3D object detection for LiDAR-camera-based autonomous vehicles using one shared voxel-based backbone,’’ *IEEE Access*, vol. 9, pp. 22080–22089, 2021.

[36] W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, ‘‘SE-SSD: Self-Ensembling Single-Stage object Detector from point cloud,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14489–14498.

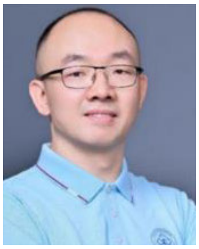
[37] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, ‘‘Digging into self-supervised monocular depth estimation,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3827–3837.

[38] J. Xu, L. Peng, H. Cheng, H. Li, W. Qian, K. Li, W. Wang, and D. Cai, ‘‘MonoNeRD: NeRF-like representations for monocular 3D object detection,’’ 2023, *arXiv:2308.09421*.

[39] M. E. Pederiva, J. M. De Martino, and A. Zimmer, ‘‘MonoNext: A 3D monocular object detection with ConvNext,’’ 2023, *arXiv:2308.00596*.

CHENTAO SUN is currently pursuing the master’s degree with the School of Physics and Telecommunication Engineering, South China Normal University. His research interests include 3D object detection for autonomous driving, key point detection, and computer vision.

CHENGRUI XU is currently pursuing the M.E. degree with South China Normal University. His research interests include monocular depth estimation and computer vision.



WENXIAO FANG received the B.S., M.S., and Ph.D. degrees in condensed-matter physics from Sun Yat-sen University, China, in 2002, 2005, and 2008, respectively. He was a Visiting Scholar with The Hong Kong University of Science and Technology, Hong Kong, in 2009. After that he joined the China Electronic Product Reliability and Environmental Testing Research Institute (CEPREI), where he has been a Research Fellow with the Science and Technology on Reliability Physics, since 2020. Since 2023, he has also been a Professor with the School of Integrated Circuit, Sun Yat-sen University. His current research interests include electromagnetic compatibility in integrated circuit and component level, and electromagnetism application in power electronics.

KUNYUAN XU received the Ph.D. degree in optical engineering from Sun Yat-sen University, in 2008. From July 2015 to July 2016, he was a Visiting Scholar with The University of Manchester, Manchester, U.K. He is currently a Professor with the School of Physics and Telecommunication, South China Normal University (SCNU). His research interests include image processing and optoelectronic technology.